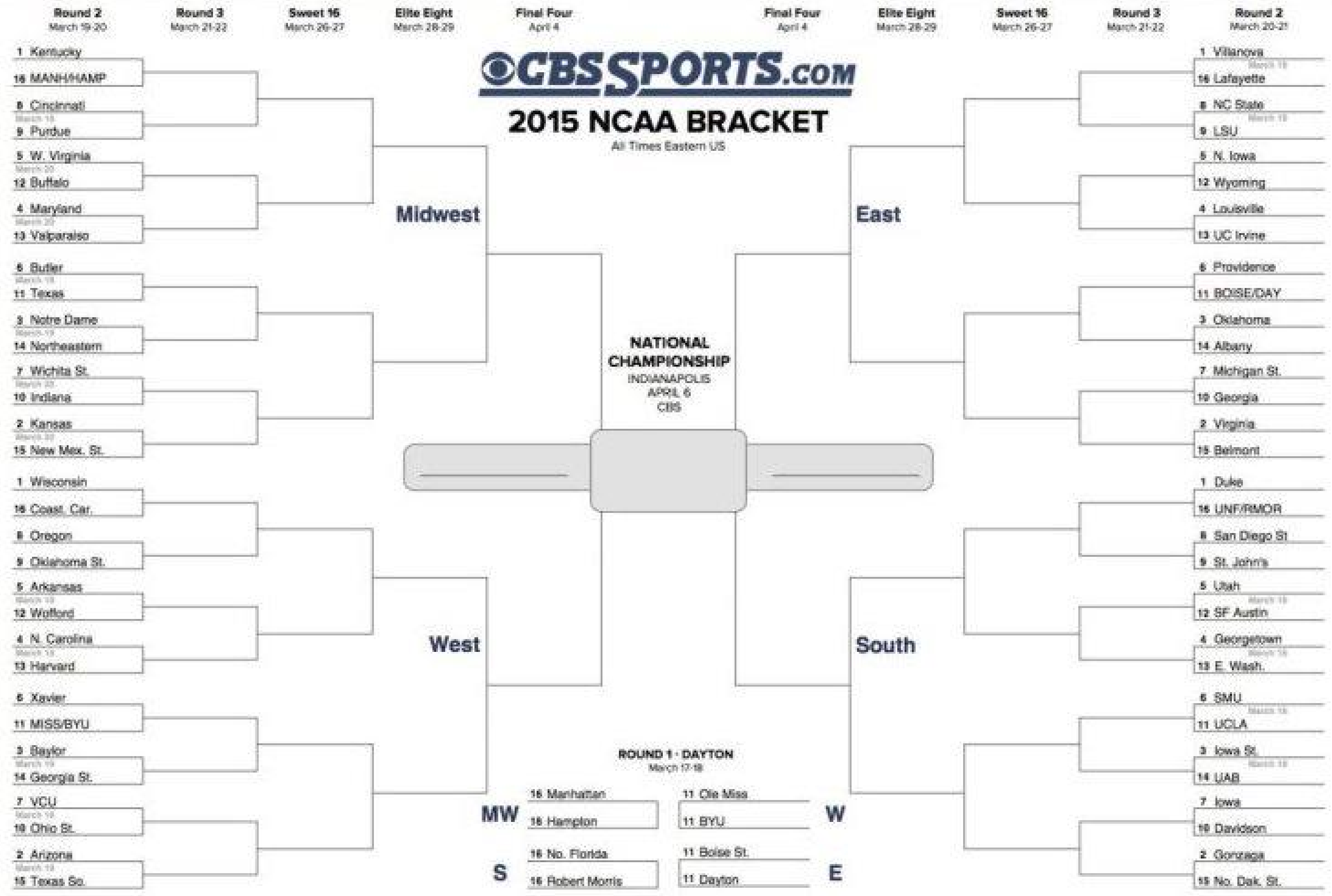
Predicting the NCAA Basketball Tournament for Fun and Profit

Three Lessons for ML Projects

```
> sessionInfo()
[1] ”June 30 - July 3, 2015”
[2] ”Aalborg, Denmark”
```

Jonathan Arfa, Data Scientist @ Magnetic
Brian Femiano, Data Engineer @ Magnetic
www.magnetic.com

What is March Madness?



CBSSPORTS.COM

2015 NCAA BRACKET

All Times Eastern US

Round 2
March 19-20

Round 3
March 21-22

Sweet 16
March 26-27

Elite Eight
March 28-29

Final Four
April 4

Final Four
April 4

Elite Eight
March 28-29

Sweet 16
March 26-27

Round 3
March 21-22

Round 2
March 20-21

Midwest

1 Kentucky
16 MANH/HAMP

8 Cincinnati
March 19
9 Purdue

5 W. Virginia
March 20
12 Buffalo

4 Maryland
March 20
13 Valparaiso

6 Butler
March 19
11 Texas

3 Notre Dame
March 19
14 Northeastern

7 Wichita St.
March 20
10 Indiana

2 Kansas
March 20
15 New Mex. St.

West

1 Wisconsin
16 Coast. Car.

8 Oregon
9 Oklahoma St.

5 Arkansas
March 19
12 Wofford

4 N. Carolina
March 19
13 Harvard

6 Xavier
11 MISS/BYU

3 Baylor
March 19
14 Georgia St.

7 VCU
March 19
10 Ohio St.

2 Arizona
March 19
15 Texas So.

East

1 Villanova
March 19
16 Lafayette

8 NC State
March 19
9 LSU

5 N. Iowa
12 Wyoming

4 Louisville
13 UC Irvine

6 Providence
11 BOISE/DAY

3 Oklahoma
14 Albany

7 Michigan St.
10 Georgia

2 Virginia
15 Belmont

South

1 Duke
16 UNF/RMOR

8 San Diego St
9 St. John's

5 Utah
March 19
12 SF Austin

4 Georgetown
March 19
13 E. Wash.

6 SMU
March 19
11 UCLA

3 Iowa St.
March 19
14 UAB

7 Iowa
10 Davidson

2 Gonzaga
15 No. Dak. St.

NATIONAL CHAMPIONSHIP
INDIANAPOLIS
APRIL 6
CBS

ROUND 1 · DAYTON
March 17-18

MW
16 Manhattan
16 Hampton

W
11 Ole Miss
11 BYU

S
16 No. Florida
16 Robert Morris

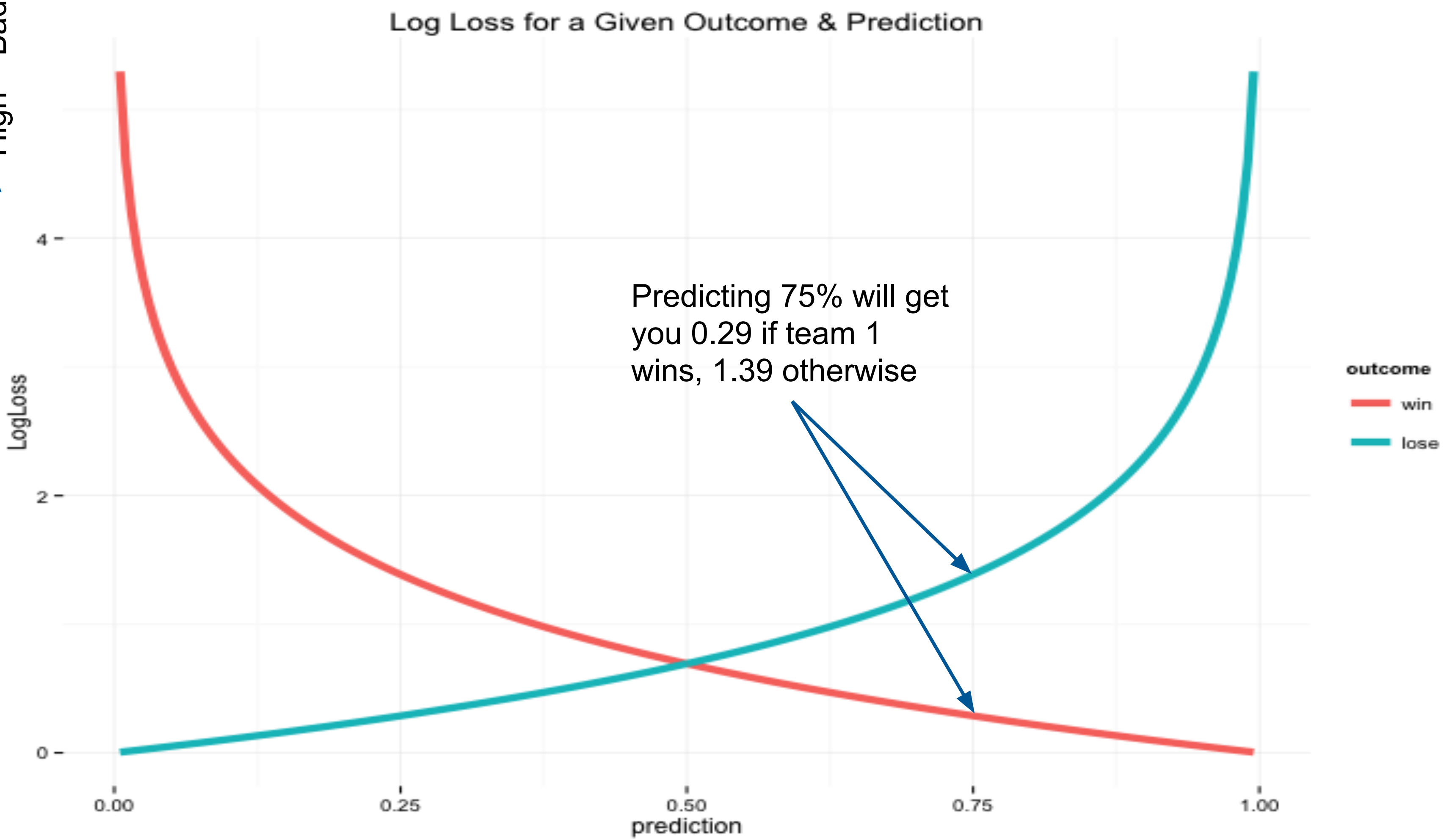
E
11 Boise St.
11 Dayton

Kaggle’s March Machine Learning Mania

- For all 2,278 *potential* matchups in the tournament, submit the probability that team1 beats team2.

- Teams judged on **Log Loss** of predicted probability (0-1) vs. actual outcome of game (0 or 1)

High = Bad



Predicting 75% will get you 0.29 if team 1 wins, 1.39 otherwise

Lesson 1: Get the Best Data

- Team-level metrics aggregated from regular season games
- Ken Pomeroy's team-level metrics (paid subscription data)
- Vegas betting odds for first-round games
- Distance traveled

Lesson 2: If Your Results Are Too Good To Be True, They’re Probably Wrong

- **Data leakage** - “the creation of unexpected additional information in the training data, allowing a model or machine learning algorithm to make unrealistically good predictions” [1]
  - your training data must represent only the knowledge that will exist when your model is run in the real world

- **FULLY UNDERSTAND AND EXPLORE YOUR DATA BEFORE USING IT**

[1] https://www.kaggle.com/wiki/Leakage

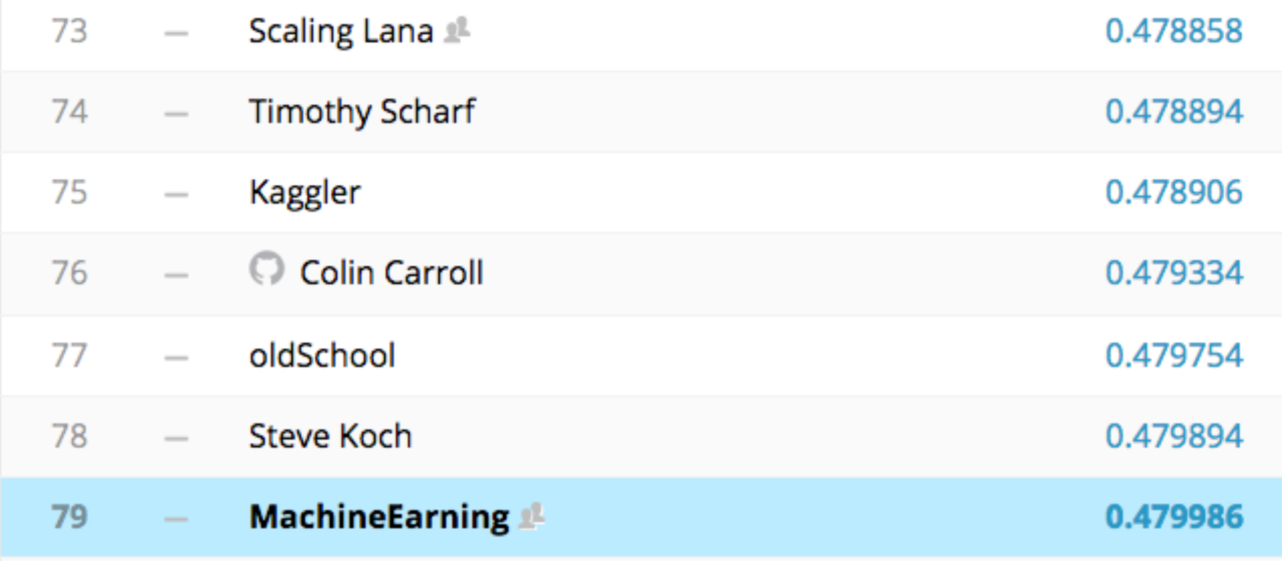
Lesson 3: Separate Yourself From the Pack

- Gamble more - manually adjust predictions for a few games.
  - Most of the top (and the bottom) teams did this.

- Unique Data / Features
  - Use the network of regular season games better
    - If team A > team B > team C, then team A > team C.

- Take a Bayesian approach to predicting later games
  - If a low-ranked team wins the 1st two rounds, it has revealed itself to be a better team than previously thought. Shouldn’t we upgrade its chances of winning the next game?

Performance



73	—	Scaling Lana	0.478858
74	—	Timothy Scharf	0.478894
75	—	Kaggler	0.478906
76	—	Colin Carroll	0.479334
77	—	oldSchool	0.479754
78	—	Steve Koch	0.479894
79	—	**MachineEarning**	**0.479986**

#79 / 341 teams

Winning Team: 0.439

MachineEarning: 0.480

Median of all teams: 0.489

https://www.kaggle.com/c/march-machine-learning-mania-2015

MAGNETIC™

R python™ Spark

Jonathan Arfa, Data Scientist
jonathan@magnetic.com
Brian Femiano, Data Engineer
brian.femiano@magnetic.com